The Impact of Explanations on Fairness in Human-AI Decision Making: Protected vs Proxy Features

NAVITA GOYAL*, University of Maryland, USA
CONNOR BAUMLER*, University of Maryland, USA
TIN NGUYEN, University of Maryland, USA
HAL DAUMÉ III, University of Maryland & Microsoft Research, USA

AI systems have been known to amplify biases in real world data. Human-AI teams have the potential to control for these biases for fairer decision-making, and there is hope that explanations can help humans understand and combat model biases. Traditionally, explanations focus on the input features that are salient to the model's predictions. If a model is biased against some protected group, explanations may include features that demonstrate this bias. However, the relationship between a proxy feature and the protected one may be less clear to a human. In this work, we consider whether explanations are sufficient to alleviate model biases due to proxy features in human-AI decision-making teams. We study the effect of the presence of protected and proxy features on participants' perception of model fairness and their ability to improve demographic parity over an AI-only model. Further, we examine how different interventions—model bias disclosure and proxy correlation disclosure—affect fairness perception and parity. We find that explanations alone are not sufficient in flagging fairness concerns when model biases are caused by a proxy feature. However, proxy correlation disclosure helps participants identify unfairness and better decide when to rely on model predictions.

Additional Key Words and Phrases: indirect biases, fairness, human-AI decision-making, machine learning, explanation

1 INTRODUCTION

Improving the fairness and trustworthiness of AI systems is often cited as a goal of explainable AI (XAI) [20, i.a.]. Research in XAI aims to improve fairness and trustworthiness by providing insights into model predictions, and thereby allowing humans to understand and correct for model biases. On the other hand, in the context of human-AI decision-making previous work has noted that humans often over-rely on AI predictions, and explanations can exacerbate this concern [5]. This is especially troubling if the underlying model contains systematic biases, which may go unnoticed even when teamed with a human. In order for the human-AI team to be successful, the human needs to be able to determine when to rely on or override potentially biased AI predictions. Although explanations can help alleviate model biases in cases when predictions are based on protected attributes directly [8, 27], these biases are often not so explicit [18, 24]. In particular, it may be difficult for humans to identify and resolve biased model predictions based on the proxy features present in real-world data, even when explanations are provided.

In this work, we conduct an empirical study considering whether explanations can help people to identify model biases and to calibrate their reliance on a model based on these biases in the context of microlending predictions.

[*] Both authors contributed equally to this research.

Authors' addresses: Navita Goyal, navita@umd.edu, University of Maryland, College Park, MD, USA; Connor Baumler, baumler@umd.edu, University of Maryland, USA; Tin Nguyen, University of Maryland, USA, tintn@umd.edu; Hal Daumé III, University of Maryland & Microsoft Research, USA, me@hal3.name.

Beyond direct bias, which is apparent through protected features, we also consider the effect of explanations when indirect bias is revealed through proxy features which may be less obvious to a human. We examine whether explicit disclosure about model biases help humans calibrate their overtrust in a biased model. In the proxy case, we further examine the role that disclosing the relationship between proxy and protected features can play in helping improve the fairness of human-AI teams.

We find that when bias is clear from the use of protected features, participants are able to identify model biases and correctly decide when to override the model's suggestions to produce predictions with higher demographic parity than the AI alone. However, in the case of indirect bias, we find that explanations alone are not enough to signal that the model is inappropriately relying on a proxy feature, let alone which examples are being unfairly penalized due to the proxy. We find that disclosing model biases—explicitly telling participants the degree of model bias—increases human-AI parity in the case of indirect biases to some extent. However, participants often fail to correct proxy-based biased predictions. When the specific correlations between the proxy and protected feature are disclosed, however, we find that participants are significantly better at correcting model biases, achieving higher demographic parity.

Our study highlights that despite the promise of explanability in fostering appropriate trust and reliance in AI systems and fair decision making, explanations alone are an insufficient tool to guide participants' trust when the underlying biases are opaque. Our results demonstrate how educating humans about potentially harmful proxies can play a vital role in helping them decide when not to rely on a biased model's predictions in order to yield fairer human-AI decisions.

2 STUDY OVERVIEW

We study the effect of explanations in improving the fairness of decisions made by human-AI teams when bias stems from different kinds of features and when participants are given different kinds of information about the model and its training data. In our study, model biases can be direct, stemming from the protected feature (gender), or indirect, stemming from a proxy feature (university) that is correlated with the protected feature. Participants may receive disclosures about the level of model bias and the strength of correlation between the proxy and protected feature. Our study addresses the following research questions:

RQ1: Does the utility of explanations in improving the fairness of human-AI teams change when models exhibit direct vs indirect bias?

RQ2: Does model bias disclosure change human-AI fairness?

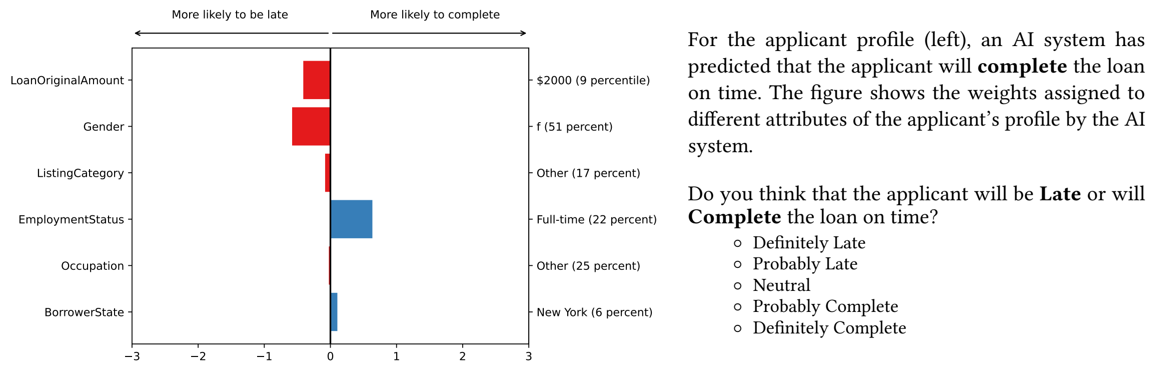
RQ3: Does proxy correlation disclosure change human-AI fairness?

We consider the utility of explanations under two lenses: the accuracy of **human's perception of fairness** and the improvement in **demographic parity in human-AI decision-making** over AI-only parity.

3 METHODOLOGY

In order to answer the research questions posed in section 2, we study decisions made by a human-AI team. In this study, the AI teammate is a classification model trained on partially-synthetic data in the context of loan prediction. We choose the task of loan prediction from a micro-lending platform as it is a decision-making task performed by laypeople which means that crowd-workers are more likely to have intuitions about the task and the features used in predictions. In our study, participants are shown either the protected feature of binary gender[1] or the proxy feature of university.

[1] We only consider binary gender in this study. Since each participant sees only a handful of examples per task phase, it would be difficult to both show the participants a statistically realistic number of non-binary applicants and get a good sense of how participants handle anti-trans model bias. We leave this for future work.



For the applicant profile (left), an AI system has predicted that the applicant will **complete** the loan on time. The figure shows the weights assigned to different attributes of the applicant's profile by the AI system.

Do you think that the applicant will be **Late** or will **Complete** the loan on time?

○ Definitely Late
○ Probably Late
○ Neutral
○ Probably Complete
○ Definitely Complete

Fig. 1. Example explanation "protected" model. The predicted outcome is completing the loan on time. The labels on the left show the name of each feature. The labels on the right show the value of each feature for the current applicant and the percent/percentile of this value in the training data. On the x-axis positive blue values correspond to "Complete" predictions and negative red to "Late".

3.1 Treatments

We examine treatments based on the directness of bias and the kind of disclosure the participant receives.

- Protected or Proxy: Whether the participant was shown a model and features including the protected (gender) or proxy (university) feature.
- Bias Disclosure: Whether the demographic parity (see section 4) of the system is disclosed to the participant. For our purposes, this is always included.
- Correlation Disclosure: Whether the specific level of correlation between university and gender is disclosed to the participant. This is never included in Protected conditions.

In this study, we consider one **Protected** condition with bias disclosure, and two Proxy conditions: one **Proxy with correlation disclosure** and one **Proxy without correlation disclosure**. We compare Protected and Proxy conditions to answer RQ1 and Proxy with and without correlation disclosure conditions to answer RQ3. In all conditions, we compare fairness before and after bias disclosure to answer RQ2.

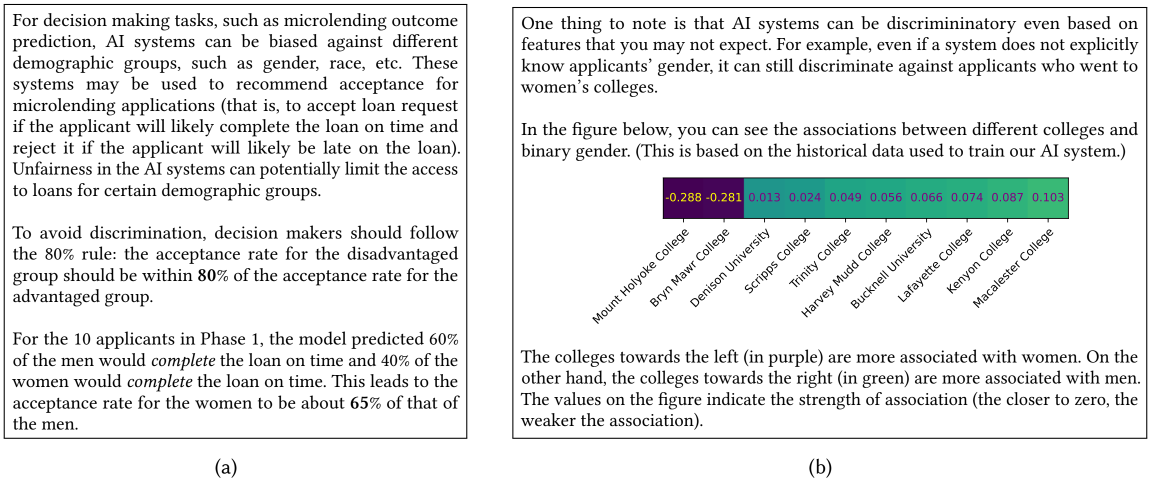
3.2 Procedure

The study procedure consists of two surveys (S1, S2), one tutorial and warm-up phase (P0), two task phases (P1, P2), and a disclosure interlude (D) ordered P0, P1, S1, D, P2, S2.[2]

Task Phases. In each task phase (P1, P2), the participant is shown 10 profiles of loan applicants: their features and explanation (Figure 1, left) as well as the overall AI prediction. They are asked to mark on a five-point-scale whether they think the applicant will complete their loan on time or be late in repaying their loan (Figure 1, right). Their response to this question serves as the decision made by the human-AI team.

In each phase, we control the distribution of gender and AI predictions. The participant sees applications from 2 women who are predicted as "Complete" and 3 women who are predicted as "Late" and vice versa for men. (This is true in the underlying data even if the participant and the model do not directly see applicants' gender.)

[2] This study design is IRB approved (#1941548-2), and participants are paid at a rate of $15/hour.



For decision making tasks, such as microlending outcome prediction, AI systems can be biased against different demographic groups, such as gender, race, etc. These systems may be used to recommend acceptance for microlending applications (that is, to accept loan request if the applicant will likely complete the loan on time and reject it if the applicant will likely be late on the loan). Unfairness in the AI systems can potentially limit the access to loans for certain demographic groups.

To avoid discrimination, decision makers should follow the 80% rule: the acceptance rate for the disadvantaged group should be within **80**% of the acceptance rate for the advantaged group.

For the 10 applicants in Phase 1, the model predicted 60% of the men would *complete* the loan on time and 40% of the women would *complete* the loan on time. This leads to the acceptance rate for the women to be about **65**% of that of the men.

(a)

One thing to note is that AI systems can be discrimininatory even based on features that you may not expect. For example, even if a system does not explicitly know applicants' gender, it can still discriminate against applicants who went to women's colleges.

In the figure below, you can see the associations between different colleges and binary gender. (This is based on the historical data used to train our AI system.)

The colleges towards the left (in purple) are more associated with women. On the other hand, the colleges towards the right (in green) are more associated with men. The values on the figure indicate the strength of association (the closer to zero, the weaker the association).

(b)

Fig. 2. a) Bias disclosure. b) Full correlation disclosure. Proxy "no correlation disclosure" conditions include the top paragraph but with the example of a hiring system using zip code + race.

To discourage participants from making decisions without any consideration of the prediction and explanation, we randomly select one application in each gender + prediction combination. After the participant has chosen their prediction on the selected profiles, we ask for a free-text explanation of why the participant agreed or disagreed with the model prediction (or was neutral). After seeing the first applicant in P1, participants are shown an attention check question, asking them to recall the previous AI prediction.

Surveys. The two surveys (S1, S2) aim to capture participants' trust in the AI system and their perception of its fairness. Both surveys include questions asking participants to rate their level of agreement with statements relating to trust (on a scale of 1-5) [16]. The surveys also include a question about the reason(s) that led to the participants' disagreement with AI such as the explanations including irrelevant features or the decisions being unfair to applicants of different genders. We additionally ask about whether the AI system was fair across different genders.

Disclosures. Before proceeding to P2, participants may be shown general explanatory materials or specific disclosures on model bias and feature correlations (See Figure 2a). While in a perfect world, such known biases could be addressed in the model itself instead of relying on human intervention, this may not always be possible. In many cases, we may have limited access to the underlying model (e.g., only having API access) or may not be able to non-superficially "debias" it [15]. In Proxy conditions, the participants are given an explanation of how models can rely on proxy features to make biased predictions, and in the "correlation disclosure" condition, participants are explicitly told the correlation between each university and gender in the model's training data (See Figure 2b).

Based on the disclosures seen, participants are asked up to two comprehension questions. All are asked whether the model's demographic parity was above 80%. Those who received correlation disclosure are asked to select one university that is highly correlated with gender.

Tutorial and Warm-up. In P0, participants are acclimatized to the task with a full tutorial example. They are shown one tutorial example with a walk-through of the task and the AI decisions and explanations. Then they are shown a warm-up example. They are first shown a version of this example with no AI prediction or explanation. This is designed to encourage participants to properly engage with the features present. Second, they are shown the same example with the AI feature explanation (still without any prediction) as this setting has been shown to benefit decision quality and support learning by encouraging participants to cognitively engage with explanations [13].

4 FAIRNESS MEASURES

Demographic Parity. We measure the demographic parity of human-AI decision making in the two task phases across conditions. We count both "Likely Complete" and "Definately Complete" as "Complete" and similarly for "Late". We discard the "Neutral" predictions from the data. Using these binary decisions, we calculate the demographic parity for human-AI teams in task phases 1 and 2 across conditions.

$$\text{Parity}(p, c) = \frac{\mathbb{E}[Y = \texttt{Complete} | \text{Gender} = \texttt{female}, \text{Phase} = p, \text{Condition} = c]}{\mathbb{E}[Y = \texttt{Complete} | \text{Gender} = \texttt{male}, \text{Phase} = p, \text{Condition} = c]}$$

where $p \in \{1, 2\}$ and $c \in \{$Protected, Proxy without correlation disclosure, Proxy with correlation disclosure$\}$ are the task phase and condition, respectively. To avoid the denominator becoming zero (when a participant rejects all male applicants in a phase), we perform Laplace Smoothing on $\text{Parity}(p, c)$. We obtain one demographic parity score for each participant's decisions in each phase. We report the mean and standard error of parity across participants.

University Parity. In the Proxy conditions, the participant does not have access to the gender feature, and therefore, the only intervention they can reasonably take to reduce bias is to predict "Complete" more frequently for applicants from women's colleges. For this reason, we also consider "university parity" in decision making, essentially calculating demographic parity with people who went to women's colleges being in the disadvantaged group and people who went to co-ed schools being in the advantaged group, regardless of gender.

Flip Percentage. Using the binarized decisions, we calculate how often the participants chose to "flip" the model prediction, i.e., how often the participant chose "Complete" when the model predicted "Late" and vice versa. We measure both general flip percentage as well as "biased" flip percentages which consider only examples in which women or applicants from women's colleges are predicted "Late".

5 SYSTEM OVERVIEW

We conduct our study using partially synthetic microlending data, with real logistic regression models producing both model predictions and explanations. Since the participant's perceptions of how the model is interacting with the profile features is key to answering our research questions, we want to avoid any potential confounding effects from using artificial or Wizard-of-Oz model explanations, or entirely synthetic data.

The scenario of predicting whether an applicant will complete microloan repayment on time or will be late is one that our participants will likely be sufficiently familiar with to have reasonable prior intuitions about what features are relevant. A challenge is that under US law, protected features like gender cannot be considered when making loan allocation decisions. For this reason, we augment our data with a synthetic "gender" feature which we correlate with outcome to induce model bias. We also generate a proxy feature, university, which allows us to finely control the level of correlation between the proxy and gender.

Data. Our loan prediction data comes from a modified set of microloans from the website Prosper.[3] The original dataset contains 79 features of microloans including their status (completed, past due, etc). We group the loan statuses,

[3] https://www.kaggle.com/datasets/yousuf28/prosper-loan

keeping the 14000 with “Complete” or “Late” statuses (with a 7:3 train-test split). This grouped loan status is the feature that the participants and the model will predict. As showing all 79 features to the participant may be overwhelming, we select 5 features that are both important to loan prediction and are likely interpretable by a layperson.

Since lenders may not “explicitly consider prohibited factors” such as race or sex,[4] we generate values for our protected characteristic: binary gender. The existing applicants were assigned a gender such that women “Complete” vs are “Late” in repayment with a 2:3 ratio and vice versa for men. This simulates historically biased data which will cause the model to associate femaleness with being late on loans and maleness with completing them.

Using the generated “gender” feature, we further generated the proxy feature: university. We include co-ed and women’s colleges, setting the joint distribution of gender and university such that most co-ed universities have relatively balanced gender ratios. For women’s colleges, the distribution reflect real-life statistics. We choose exclusively liberal arts colleges with similar US News rankings[5] to avoid confounding due to the effect of perceptions of liberal arts vs non-liberal arts schools and perceptions of school rankings.

Since, in our biased dataset, gender is correlated with outcome and, of course, the existing features are correlated with outcome, all features may be weakly correlated with gender. To confirm that university is the only strong proxy in our data, we compare the correlation of each categorical and continuous feature with gender. For continuous features (and one-hot features of each university), we use Pearson’s r coefficient. We find that the women’s colleges have at least an absolute correlation of 0.273 across Proxy conditions, whereas the maximum absolute correlation for other continuous features is 0.014, which is much lower. Similarly, for categorical features, we use Cramer’s V, finding that the university feature has at least an absolute correlation of 0.417 while the maximum absolute correlation for the remaining categorical features is 0.082, which is also lower. Overall, we see that university (especially women’s colleges) has a much stronger correlation with gender than any other feature shown to the participants.

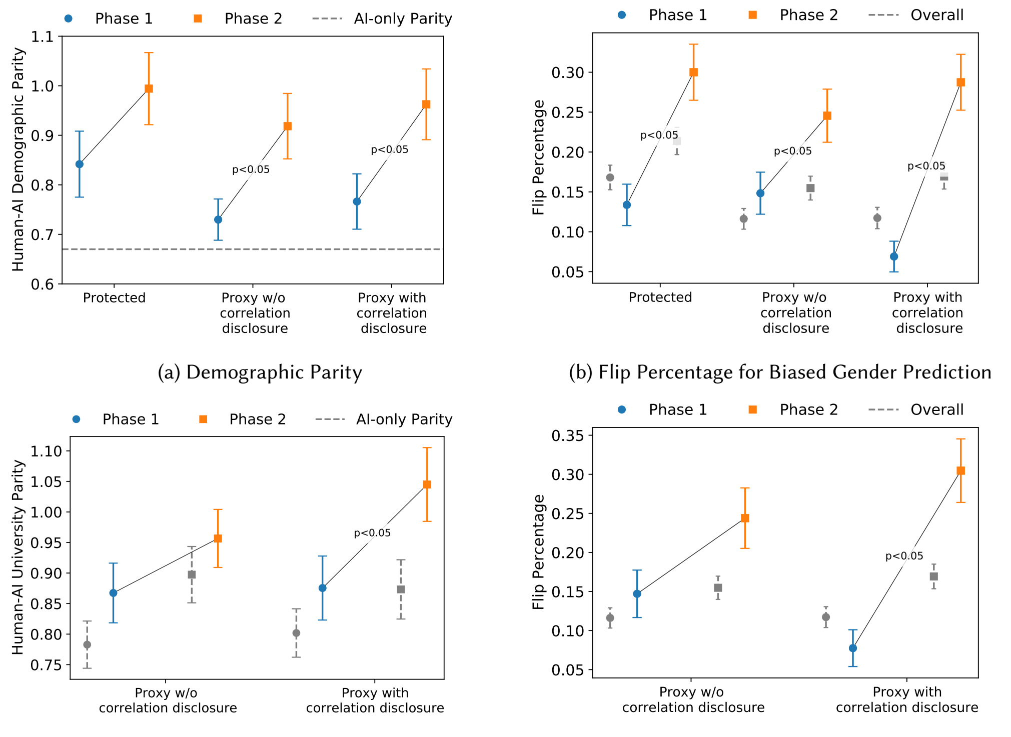
Models. For our AI predictions, we train real logistic regression models. The models are trained using 14 pre-selected features (of which participants will only see 5) and, when applicable, the gender or university feature. Since we are using logistic regression, we can create a simple input-influence explanation of the AIs’ predictions using feature weights. For continuous features like `LoanOriginalAmount`, we multiply the normalized feature value by the corresponding feature weight. For categorical features like `EmploymentStatus`, we take only the feature weight corresponding to the feature value (e.g., the weight of the `EmploymentStatus` = `Full-Time` feature). These values are graphed as in Figure 1 (left).

6 RESULTS

We recruit 70 participants per condition—Protected, Proxy with correlation disclosure, and Proxy without correlation disclosure. We discard responses that fail more than one attention check, leaving a total of 205 participants, with 68, 69, and 68 participants in the Protected condition and Proxy conditions without and with correlation disclosures, respectively. To avoid multiple testing effect, we perform Benjamini-Hochberg correction [3]. We set a false discovery threshold of 0.15 [17]. All the results are reported at a significance level of 0.05.

RQ1: Does the utility of explanation change for direct vs indirect bias? To address this research question, we study the difference in the participants’ behavior in Protected (direct bias) vs Proxy (indirect bias) conditions in Phase 1 (before any disclosure). We find that the human-AI demographic parity is higher than the AI-only parity (dashed gray line) in the Protected condition (Figure 3a). However, this is not the case in the Proxy conditions: the human-AI demographic

[4] https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_over.pdf
[5] https://www.usnews.com/best-colleges/rankings/national-liberal-arts-colleges



(a) Demographic Parity

(b) Flip Percentage for Biased Gender Prediction

(c) University Parity

(d) Flip Percentage for Biased University Prediction

Fig. 3. Fairness measures across conditions (Protected, Proxy without and with correlation disclosure) in Phase 1 (before bias and correlation disclosure, if applicable) and Phase 2 (after). The gray bar shows the AI-only parity and the overall flip rate across all predictions. The text indicates the conditions where the improvement from Phase 1 to Phase 2 is significant at $p < 0.05$.

parity is marginally higher than the AI-only parity in the Proxy conditions, but the difference is not significant. This demonstrates although useful in improving human-AI demographic parity in the direct bias case, explanations do not change the parity when the bias is indirect.

Further, as seen in Figure 3b, although the flip percentage for biased predictions (i.e., predictions of women being late on repayment) in Phase 1 is marginally higher than the overall flip percentage (dashed gray bar) for all the conditions, the difference is not significant. Thus, based on the flip percentage of biased predictions, explanations alone may be insufficient in helping participants correct biased model predictions.

RQ2: Does model bias disclosure change human-AI fairness? As seen in Figure 3a, the human-AI parity in Phase 2 improves significantly over the AI-only parity across all conditions. The improvement—the difference between the Phase 1 and Phase 2 human-AI parity—on bias disclosure is significant for Proxy conditions. This shows that model bias disclosure helps increase the parity of the human-AI team.

Further, after adding the bias disclosure (that is, in Phase 2), the rate of flipping for biased predictions is significantly higher than the baseline flip percentage (and the flip percentage in Phase 1) across Protected and Proxy conditions (Figure 3b). This result supports our finding that bias disclosure helps increase the fairness of human-AI teams.

RQ3: Does proxy correlation disclosure change human-AI fairness? As seen from the above discussion, the improvement in human-AI demographic parity and biased prediction flip rate after bias disclosure is similar for Proxy conditions with and without proxy correlation disclosure. Digging deeper into biases specific to the proxy features, we record university parity and flip rate for university-based biased predictions (i.e., predictions of people from women's colleges being late on repayment). We find that, contrary to demographic parity, the addition of correlation disclosure increases human-AI university parity over the AI-only parity (Figure 3c). This is not true for the Proxy without correlation disclosure condition. The university parity is marginally higher in Phase 2 (after adding bias disclosure) in the Proxy without correlation disclosure condition; however, the improvement is non-significant.

We observe a similar trend for flip percentage for biased model rejections for applicants from women's colleges. The difference in flip rate for all predictions vs for biased model predictions is non-significant in both Proxy conditions in Phase 1. After adding bias and correlation disclosure (that is, in Phase 2) the flip rate for biased predictions increases significantly for the Proxy with correlation disclosure condition. However, this is not true for the Proxy without correlation disclosure condition: the improvement in flip rate for biased predictions from Phase 1 to Phase 2 is marginal on model bias disclosure alone. These results indicate that beyond knowing that the model is biased, disclosure regarding the source of bias is crucial to help participants recognize and correct unfairness.

Do participants' fairness and trust ratings in the surveys reveal similar trends? The trends in participants' ratings of model fairness and trust in the model across Protected and Proxy conditions and before and after receiving bias and/or correlation disclosure are consistent with our previous findings (Figure 4). When asked whether the model was fair, participants give significantly lower ratings in the Protected condition than the Proxy conditions, even before receiving bias disclosure. This indicates that participants are able to notice fairness concerns when the biases are direct without any intervention. However, explanations alone are not sufficient to flag these issues in the case of indirect biases. We observe a significant drop in fairness ratings after adding the bias disclosure in the Proxy conditions (with or without correlation disclosure). The fairness ratings in the second survey for the Proxy conditions are still significantly higher than the Protected conditions, even though participants are explicitly told about the model biases. As these models are biased by design, we would in fact expect participants to rate these models low on trust and fairness measures.

For the trust measures, we observe worse trust ratings (high rating for wariness and low rating for the rest, e.g., safety) for the Protected condition than the Proxy conditions in Survey 1. Adding bias disclosure is not useful in the Protected condition where the fairness and trust ratings are already worse than neutral. Adding bias disclosure alone is not sufficient in the Proxy conditions as the difference in Phase 1 and Phase 2 trust ratings are only significant in the Proxy *with* correlation disclosure condition. Although we do not observe a significant effect across all trust measures, the improvement between Survey 1 and Survey 2 is consistently higher for the Proxy with correlation disclosure condition than without. This trend reflects that the bias and correlation disclosure helps calibrate participants' over-trust in AI systems when the underlying biases are indirect.

Additionally, when prompted to indicate their reason for disagreement in the survey, participants mark unfairness as a reason significantly more often in the Protected condition (48%) than Proxy conditions (3%). For the Proxy without correlation disclosure condition, the rate for selecting unfairness goes up marginally (6%) after receiving bias feedback. However, in the Proxy with correlation disclosure condition, participants mark unfairness as a reason for disagreeing with AI significantly more often (25%). This indicates that correlation disclosure is also important in calibrating participants' fairness perceptions of AI systems.

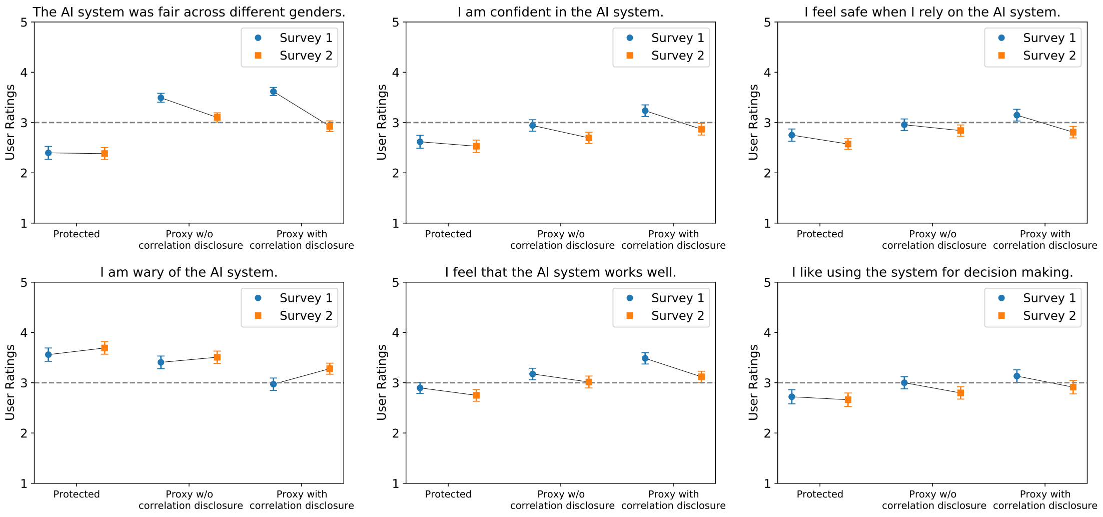


Fig. 4. Fairness and trust rating by the participants in Survey 1 and 2 across conditions. Participants are asked to rate their level of agreement regarding the statements included in the figure titles on a scale of 1 (strongly disagree) to 5 (strongly agree). We report the average and standard error of the ratings across participants in each condition. Gray bars represent the neutral line (rating of 3).

7 BACKGROUND AND RELATED WORK

Improving model fairness is often cited as a potential benefit of explainable and interpretable AI systems [2, 7, 9, 10, 20, 22, 29]. XAI is hoped to help "diagnose the reasons that lead to algorithmic discrimination" [10], to "highlight an incompleteness" in problem formalization that leads to unfairness [9], or to show compliance with fairness requirements [29].

Previous work has begun to examine how explanations affect humans' perceptions of AI systems' fairness [4, 8, 21, 25, 27]. Rader et al. [25] find that users that receive an explanation that an AI system is being used in decision-making but are not given any specific system information are significantly less likely to think the system is fair. Lee et al. [21] find that explanations of an AI system's general decision-making process do not increase perceived fairness while input-output level explanations of individual outcomes have mixed effects on fairness perception. Binns et al. [4] consider how of four different styles of explanations affect justice perception, finding no clear winner between the approaches. Dodge et al. [8] further study the explanations styles in [4]. They find that local explanations (such as presenting outcomes for similar examples) help surface fairness discrepancies between different cases while global explanations (such as describing how each feature influenced the decision for a given example) increase user confidence in their understanding of the model and enhance users' fairness perceptions. Beyond human perception of AI fairness, Schoeffer et al. [27] consider how explanations can help users appropriately rely on potentially unfair AI predictions during decision-making. They find that explanations that highlight protected features negatively affect fairness perception and that decreases in fairness perception are associated with an increase in overrides of AI predictions, even on examples where this override is detrimental to the fairness of the human-AI team.

Even without access to protected features like gender and race, models can still produce biased outcomes by relying on proxy features [18, 24]. For instance, a model that has direct access to a "race" feature and one with access to features like zip code, name, or language spoken at home could produce similarly biased predictions.

In the context of human or AI decision-making, fairness can be defined in many ways [23] with not all definitions being simultaneously satisfiable [6, 19]. Demographic parity [12, i.a.], a measure of independence between protected characteristics and prediction, has been found to be more understandable to laypeople and better capture their perception of fairness than competing metrics [26, 28].

8 DISCUSSION AND LIMITATIONS

In line with previous work [8], our findings support that explanations indeed help humans identify fairness concern when the biases are *direct*. This can be seen through participants' fairness perceptions (survey ratings), participants' behavior (flip rate for biased predictions), and participants' performance (human-AI parity). However, explanations alone are not sufficient in flagging fairness concern when model biases are *indirect*. In regard to RQ1, we conclude that the utility of explanations in human-AI teams differ for direct and indirect biases.

We further observe that model bias disclosure indeed improves participants' demographic parity and their flip rate for biased predictions as well as their fairness perception in the Proxy conditions. This reflects that bias disclosure is useful to some extent in flagging fairness concerns that go unnoticed in cases when model biases, and in turn, explanations are indirect. With regard to RQ2, we conclude that bias disclosure helps improve fairness in human-AI teams to some degree in the case of indirect bias. However, we find that participants are not able to specifically catch the biased predictions in the Proxy condition with bias disclosure alone. This can be seen through non-significant improvement in university parity and flip rate for biased rejections of applicants from women's colleges. Further adding proxy correlation disclosure shows improvement in proxy-specific biases. This indicates that specific feedback about the correlations that model rely on to make biased predictions allows participants to identify biased predictions better and correct for them. With regard to RQ3, we conclude that correlation disclosure is crucial to guide behavior to yield more fair human-AI decision making.

One caveat in our study is that our use of partially-synthetic loan data means we cannot use fairness or accuracy metrics that rely on a ground truth outcome. This, in part, led us to use demographic parity which has been argued to be insufficient as a notion of fairness [11]. We chose between two types of datasets. In settings like lending, it would be illegal to collect protected features like race and gender. These features are available in criminal risk assessment instrument datasets like COMPAS. However, this dataset is problematic and not ecologically valid [1].

Additionally, our study only included conditions where the participant is shown features, explanations, and predictions. We leave comparisons to settings where, for instance, participants are not provided explanations as future work. We also only consider input influence-based explanations; other explanation types (e.g., case-based or sensitivity based; see Dodge et al. [8]) may have different effects in the proxy case.

In this work, we study human trust and reliance in biased AI systems for a fixed level of bias and correlation between the protected and proxy attribute. Estimating human trust calibration for varying levels of model bias and correlation is left for future work. Further, we consider a singular proxy linked to a singular protected attribute. In reality, models may rely on a combination of features as proxies to one or more protected attributes [14, 24]. Our findings reveal that beyond the aggregate bias level, knowledge about specific correlation is crucial to humans for effective trust calibration. This calls for more efforts towards designing appropriate interventions for complex cases. Our work highlights the need to expand explanation beyond the influence of specific input features and incorporate underlying correlations that models may rely on.

Acknowledgements. We sincerely thank the reviewers and members of the UMD CLIP lab for their valuable feedback.

REFERENCES

[1] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/92cc227532d17e56e07902b254dfad10-Paper-round1.pdf

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[6] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047 PMID: 28632438.

[7] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *CoRR* abs/2006.11371 (2020). arXiv:2006.11371 https://arxiv.org/abs/2006.11371

[8] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310

[9] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. https://doi.org/10.48550/ARXIV.1702.08608

[10] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2021. Fairness in Deep Learning: A Computational Perspective. *IEEE Intelligent Systems* 36, 4 (2021), 25–34. https://doi.org/10.1109/MIS.2020.3000681

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[12] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[13] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

[14] Talia B. Gillis and Jann Spiess. 2019. Big Data and Discrimination. In *University of Chicago Law Review*, Vol. 86. Issue 2. https://chicagounbound.uchicago.edu/uclrev/vol86/iss2/4

[15] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 609–614. https://doi.org/10.18653/v1/N19-1061

[16] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[17] James X Hu, Hongyu Zhao, and Harrison H Zhou. 2010. False discovery rate control with groups. *J. Amer. Statist. Assoc.* 105, 491 (2010), 1215–1227.

[18] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems* 30 (2017).

[19] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[20] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[21] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 182 (nov 2019), 26 pages. https://doi.org/10.1145/3359284

[22] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (jun 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[23] Arvind Narayanan. 2018. Translation Tutorial: 21 Fairness Definitions and Their Politics. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, Vol. 1170. New York, USA, 3.

[24] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-Aware Data Mining *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

[25] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[26] Debjani Saha, Candice Schumann, Duncan C. McElfresh, John P. Dickerson, Michelle L. Mazurek, and Michael Carl Tschantz. 2020. Measuring Non-Expert Comprehension of Machine Learning Fairness Metrics. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 776, 11 pages.

[27] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *arXiv preprint arXiv:2209.11812* (2022).

[28] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[29] Richard Warner and Robert H. Sloan. 2021. Making Artificial Intelligence Transparent: Fairness and the Problem of Proxy Variables. *Criminal Justice Ethics* 40, 1 (2021), 23–39. https://doi.org/10.1080/0731129x.2021.1893932